Adding business intelligence to statistical systems
The experience of Banco de Portugal

Maria do Carmo Aguiar\textsuperscript{1}, Carlos Alberto Martins\textsuperscript{2}
\textsuperscript{1}Banco de Portugal, e-mail: Maria.Carmo.Aguiar@bportugal.pt
\textsuperscript{2}Banco de Portugal, e-mail: Carlos.Martins@eba.europa.eu

Abstract

There are several challenges regarding statistics compilation: from data collection to data processing, from data quality management to output production and dissemination. However, timely, efficient and reliable data analysis is a main challenge that has become of crucial importance in the recent years. In an environment in permanent change and in a context of financial and economic crisis like the one we have experienced, most of the traditionally available predefined analyses are not appropriate to give the adequate answers to the multiple ad hoc requests that arise every day.

Business intelligence (BI) can be described as a broad category of applications and technologies for gathering, storing, analysing, and providing access to data to help enterprise users make better business decisions. Usually targeted to large enterprises aiming to support organisation-wide analysis and integrated decision making, the new generation of BI systems comprise features that are fundamental to statistical systems, namely integration and visualisation.

In the last 10 years, Banco de Portugal has been developing and maintaining several databases based on item-by-item reporting, with proven results on reducing or eliminating previous information gaps. In particular, statistical data on securities, periodically released by Banco de Portugal, are compiled on the basis of the Securities Statistics Integrated System (SSIS) database, a security-by-security and investor-by-investor database.

This paper describes the strategy followed by Banco de Portugal towards BI which relies on the definition of an architecture framework to be followed by new statistical systems, built upon three pillars: a data warehouse, centralised reference tables and a common IT platform. The SSIS is used as an example to describe the undergoing developments concerning the evolution towards new BI architecture. The main objectives and challenges are stated and a stepwise approach is advocated.

Keywords: Micro databases, Business Intelligence, Integration

1. Motivation and starting point

There are several challenges regarding statistics compilation: from data collection to data processing, from data quality management to output production and dissemination. However, timely, efficient and reliable data analysis is a main challenge that has become
of crucial importance in the recent years. In an environment in permanent change and in a context of financial and economic crisis like the one we have been experiencing, most of the traditionally available predefined analyses are not appropriate to give the adequate answers to the multiple ad hoc requests that arise every day. Furthermore, having micro data available is not enough; there is the need for tools that enable rapid data exploration, permitting multidimensional analysis and cross-reference of multiple sources of information with different granularity.

Information technology developments led to a new trend in the way statistics are produced: traditional aggregated reporting is being replaced by item-by-item reporting. The advantages of this approach are enormous, ranging from lower reporting costs to higher compilation flexibility. Although item-by-item reporting requires dealing with larger volumes of data, this has become easier due to the evolution in network and communication protocols, database systems and multidimensional analytical systems. Consequently, old multiple heterogeneous collection and compilation systems are being replaced by integrated systems. Nevertheless, migration of existing systems to a new integrated and coordinated information system cannot be done in one step; therefore multiple sources of information must be gradually integrated in such a way that the best of each can be explored.

Banco de Portugal (BdP) statistics rely on a large set of macro and micro data information systems that are interdependent. From Balance of Payments and International Investment Position (BoP/IIP) to Monetary and Financial Institutions (MFI) and Other Financial Institutions (OFI) statistics, all of which contribute to the elaboration of National Financial Accounts.

Producing high-quality and timely monetary and financial statistics is a key responsibility of national central banks. Ensuring that these statistics remain fit-for-purpose implies keeping pace with financial innovation, assessing the statistical impact of innovations at the earliest possible stage and making the necessary amendments in a well-timed manner, if possible without overburdening the reporting agents and by making a more efficient use of the data already available.

The financial turmoil of 2007-2009 revealed important gaps in information for the purposes of financial stability analysis, namely concerning counterpart data. The development of micro-databases can have a major contribution in overcoming some of these shortcomings. They permit to develop knowledge about the activities of economic and financial agents at a more detailed level and allow for the drawing of conclusions that would not be possible should one rely solely on aggregated data.

In the last 10 years, BdP has been developing and maintaining several databases based on item-by-item reporting, with proven results on reducing or eliminating previous information gaps. In particular, it has been exploring the statistical potential of various sources of information, including the Securities Statistics Integrated System (SSIS), the Central Credit Register (CCR) and the Central Balance Sheet Database (CBSD).
The SSIS was developed by the Statistics Department of BdP with the purpose of gathering in a single repository all the information deemed necessary to comply with reporting requirements on securities. The SSIS is an information system that stores data on securities issues and portfolios (holdings) on a “security-by-security” and “investor-by-investor” basis. In addition to compiling statistics on securities, SSIS data are also used in the assessment of exposures on the balance sheet of resident financial institutions, quality control (cross-checking with other statistics) and identification of components of structured financial instruments that are not separately reported to the system. Quite ambitious in its aims, the system has been source of opportunities for data “explorers”.

The CCR database is an administrative database created to provide credit-related information to the participants for their assessment of the risks attached to extending credit. The use of CCR data for other purposes is authorized since 1996; it has been used in statistics (e.g., business register, data quality control, complementary data, separate statistical outputs); in banking supervision and regulation (e.g., assessment of credit risk and concentration of risk exposures both at micro and macro level, improvement of on-site inspection practices) and in economic research and policy (e.g., structural analysis, monetary policy).

The CBSD was, up to 2007, a database with economic and financial indicators on a significant set of Portuguese non-financial corporations, based on accounting. The reported data related to a quarterly survey, in cooperation with Statistics Portugal (mandatory); an annual survey (voluntary); and financial statements. Since 2007, data reported under the Simplified Corporate Information (SCI) replaces the CBSD annual survey. SCI is a joint electronic submission of accounting, fiscal and statistical information by companies to the Ministry of Justice, the Ministry of Finance, the Statistics Portugal, and the Banco de Portugal. It allows companies to fulfill four reporting obligations through a single submission, entirely paper-free, at one moment in each year. CBSD data is used in statistics, economic analysis, financial stability analysis, prudential supervision, research and credit risk management.

Statistical data produced at BdP is available to the public on its website. The BPstats-Statistics online, developed by BdP, is an Internet online access to relevant statistical information on the Portuguese economy (data and metadata).

Business intelligence (BI) can be described as “a broad category of applications and technologies for gathering, storing, analysing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining”.

In the last two decades, many organisations developed data warehousing projects. The scope of these developments ranged from combining multiple legacy systems to developing user interface tools for analysis and reporting. In the past, business intelligence amounted to a set of weekly or monthly reports that tended to be unconnected and were available mainly to executives, the reason why these systems were
often called executive information systems. Usually targeted to large enterprises aiming to support organisation-wide analysis and integrated decision making, the new generation of BI systems comprise features that are fundamental to statistical systems, namely integration and visualisation.

A possible strategy to deal with the issues arising from the need to fill the information gaps should not necessarily rely only on gathering new information on financial innovation-related activities and/or on restraining these activities through heavier regulation, but rather on improving the overall efficiency of the statistical framework by further exploring the largely unused statistical potential of already existing data sources, in particular linking micro-data to macro-risks. Developing a BI architecture that enables efficient data analysis can be the answer to fulfill this objective.

The remaining sections of this paper describe the strategy followed by BdP which relies on the definition of an architecture framework to be used by new or renewed statistical systems, built upon three pillars: a data warehouse, centralised reference tables and a common IT platform. The SSIS is used as an example to describe the undergoing developments concerning the evolution towards new BI architecture. The main objectives and challenges are stated and a stepwise approach is advocated.

2. Business Intelligence Architecture

Evolution of methodologies, frameworks and statistical demands are the main reasons for the need to reformulate most of the information systems that are used in statistical production. Furthermore, the need to interconnect different statistical domains and cross check data has become a fundamental requisite.

In the beginning 2008, the IT Department and the Statistics Department of BdP initiated a study with the purpose of defining a BI framework, to be used as a reference in all future IT developments in the statistical area. A layered approach was taken (largely inspired in Zachman Framework for enterprise architecture), which looked at the problem by three different perspectives, namely the business view, the information systems view and the information technology view. For each of these views there was a focus on structural and functional aspects of the problem, that were to be combined in a comprehensive and coherent vision of what we call the BI Architecture for the BdP statistical systems.

This framework is built upon three pillars: a data warehouse, centralised reference tables and a common IT platform. The data warehouse guarantees a central access point to every statistical data, independent of the input source or the production process; the centralised reference database provides common reference data and enables cross linking information from different sources and systems; the consistent usage of a common
technological infrastructure across the multiple information systems makes it easier to integrate and reuse components and promotes data access efficiency and transparency to final users.

The stocktaking exercise on the relationships between existing statistical systems showed a high level of interdependency, with multiple fluxes of data being exchanged over the network, each of which poses a need of data extraction, cleaning, transformation and integration.

As for the statistical production chain, it was noted that all systems were executing fairly comparable processes, which were however implemented in very different ways. These observations lead to the conclusion that there was a need for a much deeper revamping of statistical systems beyond the IT tooling.

A data warehousing program was proposed to deal with the overhead of multiple data exchanges. Moreover, it addressed the need for a unique repository of certified statistical micro data, which was already being claimed for some time. From now on, each new or renewed system would contribute to and use data from a single centrally managed database. This will be a continuous process as there are always new developments ongoing in the statistical area, but the goal is to get to a situation where there is a single version of trusted data, and little or no data is directly exchanged between statistical systems. The Statistical Data Warehouse (SDW) is the first pillar of the BI architecture.

The second pillar of the BI architecture is Master Data Management (MDM), which is running in parallel with systems development. This is an initiative that is being held by a specific team, for several years now, and that maintains the processes of collecting, consolidating, storing and delivering of reference data (e.g. countries, currencies, financial sectors, ...) that are used across the systems. A register of Financial Institutions has been kept for long, and recently there was an effort, still taking place, to streamline the process of gathering data from several sources and consolidating it in an historical register of all resident Companies. The master data database assigns each register an internal surrogate key that never changes; the mapping between the internal key and the corresponding external business or natural keys (different data sources often report different keys for the same data entity) are kept in the MDM system; it also keeps track of changes on business keys and other relevant attributes of the registers. This is in fact an extremely important component of the architecture, as it gives consistent meaning to the facts stored in the data warehouse, and makes much easier the crossing of fact data from different systems.
The IT platform is the third pillar. In the last quarter of 2007, the IT Department and the Statistics Department jointly carried out a Proof of Concept (PoC) for the selection of the BI Platform to be used in the forthcoming IT projects of the statistics business area. After a preliminary market analysis, and given the history of BI tools usage at BdP in the recent years, the short list was reduced to two solution providers: Microsoft and SAS. The results of this PoC showed that none of the two platforms could deliver a high level of functionality across the whole spectrum of business needs, but they were in fact complementary in the sense that the shortages of one platform were well covered by the other one. In figure 1 we see the summary of the scoring attained by the two platforms, evaluated independently on several groups of functionalities.

The final choice regarding the IT platform was a combination of two solutions: Microsoft (SQL Server 2008) for the structural components and SAS which contributes with advanced functionalities in the domain of statistical analysis and analytical workflow. Microsoft has also been chosen for the development of multidimensional models (Analysis Services), ad hoc analysis & reporting (via Excel 2007), standard reporting (Reporting Services) and dashboards (Sharepoint Server).

3. SSIS features and main components

With developments in financial markets worldwide, securities statistics have increasingly gained importance. Therefore, subjects related to coverage, quality and harmonisation of securities statistics produced in the various countries are a growing concern at the international level and, in particular, within the scope of the European System of Central Banks (ESCB). In this context, integrated statistical systems enable a more efficient and harmonised production of statistical data.
Created in 1999, the SSIS stores data on securities issues and portfolios (holdings) on a “security-by-security” (s-b-s) and “investor-by-investor” basis. It gathers in a single database detailed data on stocks and transactions of debt securities and shares and other equity (financial derivatives are not included). Both stocks and transactions are collected on a monthly basis. Information is acquired by ISIN code (International Securities Identification Number).

The existence of a reference database with individual information on securities and issuers allows for the collection of statistical information from reporting entities on a s-b-s basis. This approach implies lower reporting costs, given that there is no need for reporters to aggregate background information according to multiple criteria. Furthermore, it enables a better information monitoring and a greater flexibility when exploring data and building statistical analysis.

The reference database for domestic securities classification is managed by BdP; data are collected from several sources such as the Euronext, the Securities Market Commission (SMC), the General Government, Interbolsa, commercial databases, etc. Foreign securities are classified according to the Centralised Securities Database (CSDB) managed by the European System of Central Banks (ESCB). The aim of the CSDB is to hold complete, accurate, consistent and up-to-date information on all individual securities relevant for the statistical purposes of the ESCB.

As regards securities holdings, detailed information is collected on investments by residents in domestic and foreign securities, as well as on the holdings of non-resident investors in domestic securities. This information is reported by financial institutions, the SMC, and other resident entities.

Figure 2 illustrates the current SSIS architecture. The system relies on two relational databases and one analytical database. Collected data are stored and validated (first level of quality control) in the “transactional database”. Data estimation of missing information is also done in this database. Validated and “enriched” data are daily copied to the so-called “exploration database”. A second level of quality control is made on aggregate data, by carrying out consistency tests and comparisons with other information sources. Statistical outputs are produced from the “exploration database” as well as from the “multidimensional database”. This analytical database is a quite powerful tool since it enables user-friendly multidimensional analysis of the information.

SSIS stores information on the type of instrument, the institutional sector and the residency of the issuer/investor, prices (market value), transactions and positions associated with securities issues (issues, redemptions and outstanding amounts) and transactions and positions associated with securities portfolios (purchases, sales, stocks).

---

1 In the case of investors belonging to the households’ institutional sector, data are aggregated by the investor’s country.
SSIS promotes consistency across statistics produced by BdP. In fact, SSIS information is used as an input for the compilation of a wide set of statistics produced at Banco de Portugal: from monetary and financial statistics to balance of payments and international investment position and financial accounts. On the other hand, information from other statistical sources is used as an input for estimation of missing data and for data quality management purposes. Many of these procedures have been developed over time and are accomplished via Excel and Access tools. Integration is not always straightforward. Multidimensional analysis of securities data has been in place for the last five years and it has been a powerful tool for rapid analysis and reporting, but we can foresee many advantages in adding dimensions of analysis from other statistical systems. More than ten years after its inception it is time to redefine SSIS within the BI architecture and take even more profit from it.

4. Towards BI: work in progress

The statistical value chain comprises four phases: Acquisition, Processing, Exploration and Disclosure (Figure 3). Each phase in the value chain has a comprehensive set of core tasks. The statistical value chain guiding principles are: centralisation; harmonisation; flexibility; consistency and efficiency.

The Acquisition processes deal with collecting data from external or internal sources, validating file formats or schemas, checking and substituting business keys. When the input data is stored in the SDW, this process is very straightforward, as data is already validated and stored with the correct codes. The Acquisition database is highly volatile and keeps only transient data that is to be passed to the next phase. Where the data sources are internal systems or external entities with a specified remittance protocol, the processes will be implemented by IT specialists using MS SQL Server Integration Services. In cases when the input format is less structured and subject to change, a more agile solution must be chosen, meaning that some users will have the ability to adapt existing SAS processes to changes in external data formats.
The **Production** phase is where statistics are made, and is heavily dependent on human intervention. Quality control is done using adequate statistical techniques, followed by estimates of missing or incorrect data. Classification and aggregates calculation are also examples of processes running at this stage. The Production database stores all data relating to the current processing cycle, that is data of last period plus data from previous periods subject to revision. This database needs to be updated by users in charge of the production process and is not available to users outside its statistical domain. While core data structures shall be implemented in MS SQL Server, most processes will be built in SAS, making use of its specialization towards statistical operations, and also allowing for these processes to be adjusted by power users, whenever a different methodology must be chosen.

The **Exploration** processes are all about using data that has previously been checked or calculated by specialists in the statistical domain. Multidimensional analysis, statistical analysis, data mining, reporting and data extraction are examples of such processes. The Exploration database of each statistical system is part of the common SDW, and must comply with the same rules of organisation and naming conventions. The SDW is, by nature, intended to be accessed by a large community of users of the BdP, with different levels of read access to the stored data. As all data warehouses, this database is subject oriented, integrated, non-volatile and time variant. On top of the database, there will be SQL Server Analysis Services cubes, designed to provide an easy ad hoc access to data (via Excel), giving the user endless possibilities of data explorations, without having to know anything about data structures or query languages. SAS tools are also playing here...
an important role in data analysis and reporting, while we rely on SQL Server Reporting Services for the more standardised and corporate reports.

The Disclosure processes are focused on the statistical data dissemination obligations to external entities and also for the general public. Databases supporting these processes are always a subset (data mart) of the SDW, usually anonymised data, with a lower level of granularity and filtered by rules of confidentiality. There is, currently embedded in BdP's corporate portal, an application built to provide external access to time series and multidimensional data (BPStat). A new BdP Statistics Portal is being envisaged, and will be designed on top of MS Sharepoint.

Reformulation of SSIS according to the BI architecture will imply redefining the above mentioned processes, along with adopting the common IT infrastructure. The data model that supports SSIS needs to be redesigned in order to separate data acquisition and data processing repositories, which currently are the same. Internal reference data will be replaced with master data references, ensuring harmonisation with other statistical systems. When available, input from other systems will be provided via SDW; this will be an incremental process. Multidimensional analysis, already in place, will benefit from other systems data available via SDW. Efficiency gains are expected, along with improved data analysis.

5. Conclusions and way forward

The use of micro-databases and item-by-item reporting, covering different areas of the economy and the financial markets, has been helping National Central Banks worldwide to circumvent a number of shortcomings related to the conventional data collecting systems. The use of such data for statistical purposes can deliver significant reductions in respondent burden, higher data quality and enhanced responsiveness to ad hoc information requests from the users.

A possible strategy to deal with the issues arising from the need to fill the information gaps should not necessarily rely only on gathering new information on financial innovation-related activities and/or on restraining these activities through heavier regulation, but rather on improving the overall efficiency of the statistical framework by further exploring the largely unused statistical potential of already existing data sources.

The strategy followed by Banco de Portugal relies on the definition of an architecture framework to be used as a reference in all IT developments in the statistical area. This architecture is built upon three pillars: a data warehouse, centralised reference tables and a common IT platform.

Following a stepwise approach, the implementation of this architecture framework will contribute to the construction of a coherent and truly integrated statistical system as opposed to having multiple systems that coexist but are not connected in an efficient way. At the moment, three statistical information systems are being reformulated according to this model: the BoP/IIP, the CBSD and SSIS. The outcome of this effort will soon be evaluated.
References

Banco de Portugal (2008), Supplements to the Statistical Bulletin: 2/2008 “Securities Statistics – Integrated System Features and Main Results”